Northumbria Research Link

Citation: Foster, Jonathan, McLeod, Julie, Nolin, Jan and Greifeneder, Elke (2018) Data Work in Context: Value, Risks and Governance. Journal of the Association for Information Science and Technology, 69 (12). pp. 1414-1427. ISSN 2330-1635

Published by: Wiley

URL: https://doi.org/10.1002/asi.24105 <https://doi.org/10.1002/asi.24105>

Data Work in Context: Value, Risks and Governance

Jonathan Foster, Julie McLeod, Jan Nolin, and Elke Greifeneder

Abstract

While always integral to scientific activity, data work has recently emerged as a key set of processes within societal activities of all kinds. While data work presents new opportunities for discovery, value creation, and decision-making, its emergence also raises significant ethical issues, including those of ownership, privacy, and trust. This article presents a review of data work, and how negotiating a trade-off between its value and risks requires locating its processes within the contexts of its conditions and consequences. These include international, national, and sectoral conditions of law, policy and regulation at a macro level; organizational conditions of information and data governance that aim to address the value and risks of data work at a meso level; along with attention to the everyday contexts of data and information handling by data information and other professionals at a micro level. In conclusion, a conceptual framework is presented which locates the processes of data work within the matrix of its macro meso and micro conditions, its consequences for individuals organizations and society, and the relations between them. Suggestions are given for how research into the study of data work—its value risks and governance— can be advanced by using this framework.

Introduction

The organized process of capturing, organizing, analysing, and using data has always been integral to scientific activity. Its use is key to the exercise of the scientific method and the generation of reliable knowledge. While remaining key to scientific activity, the turning of data into action, or data work, is a set of processes that can now be regarded as integral to societal activities of all kinds (Foster & Clough, 2018; Foster, 2016; Taylor, 1986). Due in large part to the development of a pervasive infrastructure of networks, smartphones, sensors and other devices, new data-driven interactions between people and organisations have emerged. Search is ubiquitous, the gathering, analysis and use of data by social media companies is routine, health providers engage in clinical data sharing for the purposes of conducting secondary research analyses and improving treatment outcomes (Lea et al, 2016; Willbanks & Friend, 2016), governments make use of data analytics to inform the making of law enforcement decisions (Bachner, Ginsberg, & Hill, 2017), while business organizations draw on data analytics in order to improve their internal operations and to inform the development of automated data-driven services (Davenport & Harris, 2017).

While the value and uses of data work are readily apparent, there is also a widespread recognition that its processes also give rise to different categories of risk. These include detectable risks to people and to organisations, along with less detectable societal risks. The risks to people principally arise from how organizations' processing of personally-identifying information may contribute to loss of privacy, via a lack of informed consent or lack of respect for client confidentiality for example, via re-identification related to data linking and data sharing, or via identity theft and loss of reputation due to security breaches. The risks to organisations arise from a lack of systematic attention to data work, e.g. lack of organisational policy and governance, and hence a loss of value and opportunity arising from its implementation; or from not adhering to regulatory and other external standards, e.g. data protection law, and sector-specific regulations that can lead to mis-handling and data breaches. Finally, there are the more amorphous and less detectable societal risks, e.g. bias, unfairness and manipulation that can arise in relation to algorithmic decision-making (Dormehl, 2014).

It is clear that the utility of data as a resource and its handling by agents of all kinds, typically but not only data and other professionals, is dependent on negotiating the trade-off between the value and risks of data work. The addressing of this trade-off at the micro level of data-handling, interaction, and use has steered attention to the macro and meso level conditions that shape and constrain data work and its processes at a micro-level. These conditions include attention to the ethical questions that arise in data work, laws, policies and other sectoral standards that inform and regulate organizations' handling of data, the systematization of organizations' data handling practices and their governance, along with attention to the implications for educating and training data professionals and others who handle data on a routine basis.

The purpose of this article is to provide a synthetic review of the value, risks and governance of data work. The review is organized as follows. An initial section on 'data' defines the scope and use of the term as used in the review. This is followed by a section on 'data work', initially addressing its organization and sub-processes, before providing illustrative examples of its values and risks from two principal perspectives: health and business (Taylor, 1986). In keeping with these perspectives, primary consideration is given to contexts of data work outside of the immediate domain of science. We do not include attention to other pertinent types of data work, e.g. data archiving, data curation, where issues of ethics, intellectual property will also arise. The section on data work is then followed by a series of sections that address macro and meso level conditions that will influence a trade-off between the value and risks of data work at a micro-level. This includes attention to ethical and legal questions, e.g. issues of ownership and intellectual property rights, privacy and data protection law at the macro level; issues of governance and accountability within organizations at a meso level; plus issues relating to the education and training of data professionals and others that inform the handling of data at the micro-level. While topics relevant to the value, risks and governance of data have been studied separately (Bacher, Ginsberg, & Hill, 2017; Lane, Stodden, Bender, & Nissenbaum 2014; Martin & Shilton, 2016; Shilton, 2012), their interdependence and interaction within the context of data work has been accorded less attention. From this emerges an integrative framework, and related set of research questions, for studying and researching the challenges that arise from data work, the conditions impinging on data work, and the consequences for individuals organizations and society.

Data: An informational interpretation

The term 'data' has been used in many different senses. These senses include: data as a given, data as the premises on which a subsequent calculation may be made, data as the evidence input into a problem, or data as an annotation – the latter a sense akin to what is now more commonly termed metadata (Furner, 2016). Of the many senses of the term, 'data' is used here in the sense of "content…about a referent" (Furner, 2016). In other words, data is used in an 'informational' sense to indicate content about an entity e.g. person, object, action, spatial location etc. It is a corollary of this definition that content about the entity will also vary along a range of attribute-values.

For the purposes of this review, this informational sense of data subsumes other ways of considering data. This includes the distinction between data in either numerical or non-numerical, textual, form —a distinction that will be pertinent to an organization's mapping of its data resources for example, but relevant here only to the extent that such data are to be interpreted as carrying actual or potential content about a referent; as well the key distinction between small and big data (Borgman, 2015). While big data clearly act as a contemporary condition for data work, it is the complexities of its volume, variety, and velocity and how

these impact on the informational value, risks, and governance of its processing that is of most relevance in the context of data work. A final distinction, that between data and digital data, also merits consideration. While this is also a useful distinction, for observing the contemporary significance of data in its predominantly digital rather than analog form—i.e. numerical or non-numerical data that is computable and available to be processed in machine-readable form - the reduction of data to digital 1s and 0s is again considered pertinent here only to the extent that digital data carries actual or potential informational content about a referent. In other words, how digital data contribute for example to the value, (e.g. real-time processing), risks (e.g. algorithmic decision-making), and governance of data-driven services that carry content about people, location, and other attributes. It is also worth bearing in mind that it in this informational sense of data where a good many of the issues surrounding the value, risks and governance of data work lie. Since apart from the integrity of data work as part of the scientific method, it is how data point either directly or indirectly to a subject, (e.g. citizen, patient, consumer, user) that leads to much of the value (e.g. personalisation), risks (e.g. loss of privacy), and governance challenges (e.g. accountability, data protection) that arise. Finally, the informational sense of data enables a further distinction to be made between data consisting of content about a referent, and the set of processes that are subsequently applied to that content—processes potentially enabling the systematic transformation of data into information, knowledge and action (Taylor, 1986). A topic to which we now turn.

Data Work

A number of frameworks have been devised for systematizing the process of capturing, organizing, analysing, and using data. These include: a value chain (Miller & Mork, 2013), an analytics value chain (Stein, 2012), and a data value cycle (OECD, 2015); while Carter and Sholler (2016) refer to 'data science work' and 'data analysis work' as a "kind of work…being done 'on the ground'" by data analysts who extract, analyse and use data in support of organizational goals. In this review, data work is conceived as a set of processes organized in accordance with Taylor's (1986) value-added approach (see also Foster and Clough, 2018; Foster, 2016). A number of reasons make Taylor's approach an apt choice. First, the value-added model begins with data, and then proceeds via information and knowledge to action (Taylor, 1986: 6). The definition of data that Taylor provides is also intentionally general, referring to 'symbols that designate the state of an entity at some point in time'. Therefore this definition therefore provides a useful anchor, one compatible not only with an informational interpretation of data, but also sufficiently flexible in order to accommodate new characteristics of data in the contemporary era. Second, Taylor makes a useful distinction between processes that leave the 'data' unchanged from input to output, and processes that do not. For example, while an abstracting and indexing service will add value to the organisation of documents via indexing processes that provide intellectual access to these documents, the underlying 'data' will remain the same. In other types of 'data work', the set of processes applied to the data, will transform that data, and the output will contain a "substantive difference" from what was input. Taylor labels these latter processes "information analysis" services. In sum, and as the distinction implies, Taylor highlights a key difference between 'data', and the processes applied to this data. This distinction is maintained here, and the value-adding 'organizing', 'analysing', 'judgmental', and 'decision' processes that can be applied to 'data' are called data work. Third, in the course of drawing attention to the transforming nature of 'information analysis services', Taylor also makes an additional observation as to the starting point for these analyses. In principle, the starting-point for analysis can be problem-oriented, e.g. when providing data analyses for example in

support of the resolution of clinical problems. In a different setting, such as business, the starting-point for analysis can be more discovery and data-driven, seeking to detect patterns that are tied more to questions, rather than to the resolution of problems. The latter is more pertinent to automated data-driven services. This distinction is also maintained here via the use of illustrative examples of data work, from a problem-oriented health domain on the one hand, and an automated data-driven business domain on the other. Fourth, Taylor's value-added approach also draws attention to the key process of decision-making—a key outcome of data work, irrespective of whether information analyses are used to inform processes of either human or automated decision-making.

Data work: value and risks

As a way of specifying and organizing data work, and in keeping with Taylor's value-added approach, we follow his broad set of 'organizing', 'analysing', 'judgmental', and 'decision' processes. We maintain this set of distinct sub-processes in general only, and do not go into detail as to the values that can be added in data work. This is a question for future research.

Organizing processes: *value and risks*

In a health setting, the process of adding value to data in health begins with the organized collection of a number of kinds of data, including data about patients in support of the resolution of clinical problems, data about processes of health administration, and data about financial transactions. For example clinical data will consist of readings from remote sensors, meters, and other vital sign devices; biometric data including fingerprints, genetics, handwriting, retinal scans; x-ray and other medical images; blood pressure, pulse and pulse oximetry readings; omics data; along with the collection of further structured and unstructured clinical data about patients, e.g. electronic medical records, physicians' notes, email, and paper documents (IHTT, 2013; Costa, 2014; Ragupathi & Ragupathi, 2014). Health administrative data will include human resource data that enables the monitoring and auditing of organizational performance (Faulds et al., 2016)–the latter incorporating data on treatments performed and targets met; while financial data about transactions, and health care claims will also be processed (IHTT, 2013). Besides the well-documented risks arising from the collection, and subsequent processing, of personally-identifying information, the main risks of organizing data from the health provider's perspective, relate to the opportunities of organizing big data. In other words, the loss of opportunity and risks to value of unified data-driven health services not being developed due to the quality, variability, and veracity of the data collected (Janke et al., 2016; Kambatla et al., 2014). In addition, there will be resistance to the development of such unified data-driven services (Costa, 2016) from patients and others (Janke et al., 2016), citing privacy and other concerns.

In a business setting, the process of adding value to data for the delivery of automated services will incorporate the collection and organization of a range of different types of personally-identifying information, including: account information, site or application interaction history, precise location-based information; along with other interactions both prior to and during interaction with the service, for example third-party referrals, site navigation. Data about what information users share, with whom, via which channels (e.g. instant messaging, social media, e-mail) will also be gathered. In this way, a profile of the actions and behaviours of an individual user can be stored and exploited for the subsequent delivery of targeted content and advertising. For the purposes of gaining further insight, data-driven services will also seek to automate the collection and subsequent processing of information about their users' actions and behaviours on other connected platforms. In this way the automated and implicit capture, pooling, analysis and display of PII, can be supplemented by the gathering of explicit user-generated content on social media channels. In

contrast to the risks to value arising from the unified and organized collection and subsequent processing of big data, the risks of collecting and organizing data for automated data-driven services will primarily relate to user concerns surrounding the extent, use and onward processing of personally-identifying information.

Analyzing and judging processes: value and risks

In a health setting, the processes of adding value to data at the analysing and judging phases of data work incorporate a number of information analyses based on the application of data analytics to the types of different data collected, e.g. clinical, administrative, and financial. This includes the application of analytics to clinical audio data, which have been used to analyse a patient's communication patterns or to make judgments about their emotional status; along with the use of predictive analytics that can "target identification of early readmissions risk", and "facilitate population health management and value-based accountable care", the latter of which has also emerged as a core strategy for detecting fraud in claims for health insurance (Edelstein, 2013:16). The application of analytics to administrative data can also support the preparation of reports on, and release of, datasets to external bodies and government related to health performance outcomes.

The risks of information analyses in health relate to issues of: propensity arising from big data analyses that may pre-judge the likelihood of a certain outcome, for instance survival rates; issues of data ownership, governance, and standards including data sharing with external organisations; physicians' views on the costs, risks and liabilities of working with data (Neff, 2013); issues of trust and ethics (Childs & McLeod, 2015; McLeod & Childs, 2018), plus patients' expectations surrounding privacy in the face of demonstrable empirical evidence of continuing data breaches (Kambatla, Kollias, Kumar, & Grama, 2014; Kayyali, Knott & Van Kuiken, 2013).

In a business setting, the processes of adding value to data as part of information analyses in support of automated data-driven services include the use of sentiment analysis, summarization, and other content-analytical techniques for the analysis of textual data; the analysis of audio data, (e.g. recorded and live calls to call centres) via transcription, indexing and searching of speech content; and the analysis of video data via indexing and searching of video content (Gandomi & Haider, 2015). Text, audio, and video content-related techniques can also be used in conjunction with further network structure-based techniques for the purposes of extracting other participants and relations, identifying implicit communities, and modelling and analysing social influence; while predictive analytics aimed at anticipating future customer behaviours can also be applied to each of these same data types (Gandomi and Haider, 2015).

The risks of information analyses in support of automated data-driven services primarily relate to transparency, trust in big data analyses, and the manipulation of reputation. Indeed, these are concerns of all data mining technologies (Pang and Lee, 2008); as they are of the algorithmic processes, of which data mining forms an element (Dormehl, 2015). In addition to the risks directly arising from the process of information analysis, the key risk remains the quality of the data upon which the analyses are based. Since big data analyses rely less on deductive-driven statistical analysis and more on the inductive discovery of patterns and correlations, the heterogeneity of data, accumulation of noise, spurious correlations, incidental endogeneity and other measurement errors also represent specific challenges to the use of big data analytical techniques. Such risks also represent a risk to the acceptance of big data analyses and a firm's perceptions of the validity of such analyses (Kwon, Lee & Shin, 2014).

Decision processes: *value and risks*

In a health setting, the resulting information analyses will both inform and provide productive knowledge in support of clinical decision-making about personalized patient care; as well as administrative decision-making relating to the effective yet economic delivery of healthcare. The latter will include the return of performance data to external authorities that can in turn be used to inform decision-making in relation to managing a population's health. In summary "…big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions" (Ragupathi & Ragupathi, 2014: 1).

The risks of data-driven decision-making in health, include concerns not only about the integrity and validity of (big data) analyses but also perceptual risks. Clinical professionals for example may require considerable persuasion as to the worth of investing in (big) data analytics and its relevance to clinical priorities; and how the benefits of this data work outweigh the costs, risks, and liabilities involved (Neff, 2013).

Table 1: *Data work: value and risks*

Value	**Risks**
Organizing processes	
Organization of heterogeneous data streams	Loss of opportunity and risk to value due to insufficient exploitation of data as a resource
Unified data models	Value not realized due to the quality, variability and veracity of the data captured
Development of data-driven services	User concerns about the unwarranted processing and re-use of personally-identifying information
Analysing and judging processes	
Electronic data storage	Data breaches
Descriptive analyses across a range of data types	Personal identification and re-identification
Predictive analyses of preferences and behaviours	Measurement errors; transparency and trust
Data sharing	Issues of ownership and privacy
Decision making processes	
Economic value	Costs and benefits e.g. intuitive vs. data-driven decision making; organizational capacity for making data-driven decisions vs. other organizational priorities
Organisational performance and accountability	Administrative costs
Recommendations and micro-targeting	User manipulation
Automated decision-making	Algorithmic bias

In a business setting, the values of automated data-driven decision-making are several. First user and use value oriented, in which descriptive data analytics provide users with the additional value of real-time information in order to inform different kinds of decision-making, e.g. travel planning (TomTom, FlightStats), financial decisions (Xignite), or leisure bookings, (OpenTable). These descriptive data analytics can be supplemented with the use of

predictive analytics, and network structure-based techniques, the latter of which is used to inform social media relations. Second, provider- and exchange value oriented. The analysis of closed circuit television footage in retail can lead to insights for example into customers' queuing behaviour and how customers move around a store: "Valuable insights can be obtained by correlating this information with customer demographics to drive decisions for product placement, price, assortment optimization, promotion design, cross-selling, layout optimization, and staffing" (Gandomi & Haider, 2015: 141). Finally, with the advent of data-driven interactions, the value of automated data-driven services also rests in their capacity to provide precisely tailored content and recommendations; plus via the capture and analysis of further data prior to, during, and after the immediate service interaction, maintain a continuous relationship with their users. This set of processes, along with an illustration of their use in different problem-oriented and data-driven settings, provide an initial illustration of the value and risks of data work at the micro-level. Table 1 provides a summary of the value and risks identified.

Data Work in Context

The conduct of data work entails attention to negotiating a trade-off between the value and risks that its processing presents. The addressing of this trade-off requires locating data work within the context of its conditions and consequences. These conditions exist at a macro (e.g. international, national, sectoral), meso (e.g. organizational, institutional, consortium-based), and micro level (e.g. education and training of data professionals and others involved in systematic data-handling).

Data work at a macro level

With regard to the macro conditions of data work, the literature comes from several research areas: law, information ethics, computer science, cultural policy, information policy and mass communication. For the purposes of this review, we draw on this literature and focus on the structural conditions pertaining to two issues relevant to the value and risks of organizing and analysing data at a micro-level: the history, regulations and policy surrounding data ownership; and the structural conditions relating to the value and risks involved in the processing and re-use of personally-identifiable information. This latter is especially pertinent in an era when a contextual understanding of privacy is becoming both more key and prevalent for users and providers alike.

Ownership actually involves at least four separate sub-issues: that of the service user, the content rights holder, the service/distribution owner and the third-party data-driven innovation process owner. While the use of personal data has always been protected in its use for trade (OECD, 1980; 2011), the re-use of personal data, and the advent of big data unavoidably increase the vulnerabilities between endpoints, and both legal and criminal exploitation must be considered. These issues overlap and entangle with each other and involve research reflecting on two separate policy streams: intellectual property rights (IPR) and data protection. Both issues have emerged from doctrines on personality rights (Bygrave, 2002). Notions of copyright have been influential in the grounding of privacy rights and vice versa.

Data ownership

Historically, IPR has for several centuries played out as a competition between two different views (Sell & May 2001). The first view is weak ownership for the content rights holder, maintaining that the public should be allowed free access to information as this is a prerequisite for the development of culture, society and democracy. As an extension, this ensures an educated labor force, increases in entrepreneurial activity and economic growth.

With the second view, strong ownership for the content rights holder, the argument is made that creators of original content should be well compensated. As an extension, this creates drivers for innovation, creativity and entrepreneurship.

The Global Agreement on Trade in Tariffs (GATT), created in 1947, had been a driver of international legislation on intellectual property rights since the 1970s. When GATT morphed into the World Trade Organization (WTO, 1994) in 1995, it was constituted with three main agreements that all members were obliged to follow (Crews, 1998; May, 2003). One of these was the Agreement on Trade Related Aspects of Intellectual Property Rights (TRIPs). The TRIPs agreement signaled a global switch from a weak to a strong ownership viewpoint and therefore constitutes a crucial foundation for data-driven innovation. It is in this rather simplified polarity between producers and customers of culture, that behavioral tracking through sensors and Internet usage is introduced. The legal status of who owns the data generated by peoples' online and off-line every day practices has until today not been clarified.

It is notable that the turn to a strong ownership viewpoint served as a disturbance in what was otherwise an emerging zeitgeist of weak ownership, i.e. privileging the user of the service. However, as practices of filesharing were seen to threaten industrial interests of culture and software, it was deemed necessary to develop legal frameworks that allowed for tracking of Internet behavior. Perhaps unintentionally, the TRIPs agreement created such a foundation for national legislation on digital IPR. The focus of the agreement was actually producers and sellers of pirated material goods. An additional problem with IPR legislation is that notions of ownership from an analog context have been applied to the digital marketplace without any concern about materiality and immateriality. IPR legislation has been called "path dependent" (Litman, 2001) as the legal framework concerning illegal copying of material goods also came to include digital goods. This created a legal framework within which ordinary users were seen as upholding the same criminal status as that of professional distributors of pirated brands simply by virtue of copying a musical track, a piece of software or any other kind of digital product (Larsson & Svensson, 2010).

The most influential national policy stream developed through TRIPs was the 1998 US Digital Millennium Copyright Act (DMCA, 1998) that emphasized Digital Rights Management (DRM). Four new paragraphs were introduced into title 17, Chapter 12 of the US Code (U.S. Congress, 1998). The most important issue was how to deal with the entity "digital copy" (Gillespie, 2007). Two main ideas were introduced that would be of tremendous importance for the development of future digital practices. First, in paragraph 1201: 'Circumvention of copyright protection systems', removal of any kind of copyright protection system was prohibited. This set the stage for DRM. Second, there was an important shift in the rights of the user from ownership to leasing. This involved an emphasis on the end user license agreement which allowed a market shift from buying to license-to-use. License-to-use could, for instance, involve the right to play music legally acquired on a CD player but not on a computer. As a consequence, music freely downloaded from the file sharing platform Napster, started in 1999, appeared to be a superior product as MP3-files did not contain any DRM protection system and could be played on any device (Smith, 2003; Sterne, 2006). Given the ease of Internet-based file sharing, the enforcement of DRM required development of systematic surveillance technologies of digital watermarking and registration of who does what (Bygrave, 2002). However, such practices also raised concerns relating to privacy and data protection.

At an ambitious Policy Colloquium co-convened by Google and O'Reilly Media (Hemerly, 2012), it was suggested that harvesting data from the private sector, the public sector, and from research required separate policy perspectives. Notions of licensing data were probed in order to create a stronger sense of ownership compared to that given by

contractual obligations. However, the consensus seemed to be that extended scope of licenses would, in today's policy context, be difficult to establish. Similar problems were connected to the notion of data commons as people often are not good at anticipating usages of the data they make available. The copyright management system of YouTube, Content ID, was taken as an example of a technology that balanced the rights of the user with the appropriate data access needed for innovation. Hemerly (2013: 28) has argued that in some cases "regulation that intends to protect individuals constrains the use" and that, therefore, regulators need to create structures for voluntary informed consent participation.

Essentially, the strong ownership viewpoint of the rights holder has allowed the development of numerous data practices. Still, conventional contractual agreements seem to be inefficient for numerous usages. At the same time it is prudent to take note of the observation by Sell and May (2001) that the balance between strong and weak ownership historically has shifted on a regular basis. Given the revelations by Edward Snowden, starting in 2012, (Greenwald, 2014) it is quite possible that policy is beginning to move toward another shift.

The new European General Data Protection Regulation (GDPR) in 2016 (European Parliament and Council, 2016) introduces specific digital rights for European citizens including stronger opt out rights, the right to be forgotten and rights to access data stored about them. This regulation also broadens and makes more distinct what is meant with *personal data*, clearly specifying email addresses, websites, medical information IP addresses, cookies, genetic data, biometric data, fingerprint data, retinal scan data and location data. In addition, the traditional view of what is considered *sensitive* personal data has also been broadened. The traditional list included racial or ethnic origin, political opinions, trade union membership, religious or philosophical beliefs, health data and sexual orientation. GDPR adds to this list genetic data and biometric data. This is a radical change with fundamental repercussions for data work.

Data work and changing contexts of privacy

The end user license agreement was used as a basis for a contractual model, regulating data exchange, involving two other key documents: terms of service and privacy policy. These documents emphasized the traditional problem of content rights holder and customer. Typically, users of social media platforms such as Facebook or Google+ were seen as content rights holders of the documents, music, videos etc. produced. More importantly, behavioral data, which lacked regulatory shielding, was collected, processed, packaged and sold in various ways (Gehl, 2013). Collection, processing and packaging of data was in this way developed without a firm legal framework defending the rights of users. In the absence of legal privacy requirements, established practices of de-identification, including anonymization, key-coding, encryption, etc. were established. In addition, Privacy-Enhancing Technologies (PET), in which users control and manage their individual digital identities were developed (Hansen et al., 2004). Nonetheless, critics have argued that such practices or tools are insufficient in supplying credible protection in the face of existing and evolving tools for re-identification (Ohm, 2010).

The development of privacy by contract and the surveillance of DRM access became connected to the business model of web 2.0, as suggested by O'Reilly (2007). This built on the strategy of generating massive user generated content on free-for-use proprietary platforms and thereafter packaging and selling behavioural data. Following this, social media platform owners have explored potential forms of commodification in partnership with an evolving business sector of data brokers. These rapidly developing practices have flourished despite a rather fragile legal context focused on IPR, i.e. user generated content, rather than user generated behavior.

Through contractual agreements, social media users now routinely accept that platform owners either exploit behavioural data or the data being produced or both. Given that the most powerful corporate actors are US-based (such as Google, Amazon, Facebook, Apple and Microsoft) it is American legal requirements that have served as a norm for the development of regulation in most countries. The most important restriction placed upon platform owners in the US, and therefore for the rest of the world, is the necessity of allowing users the opportunity to "opt out", i.e. rejecting the routinized transfer of ownership and selling of personal data. However, Turow (2012) found that available functions for generating individual consent in the form of opting out often were difficult to find, navigate or simply not working as advertised. Tene and Polonetsky (2013) considered the development of a practice of "opt in" rather than "opt out" as a possible way forward. Another approach, building on the agency of users, has been to suggest "privacy by design" which involves the building of databases without any personal data (Hustinx, 2010). This strategy is closely connected to PET as well as to the principle of data minimization, which involves minimized amount of personal information stored and shared as well as duration of time stored. The emergence of metadata capital on the part of providers (Greenberg, Ogletree, Murillo, Caruso & Huang, 2014; Greenberg, 2014) along with practices of participatory personal data on the part of users (Shilton, 2012), continue to make this aspect of platform use a matter of live debate. Notably, many of these ideas, including overarching principles of privacy by design have been integrated into the European GDPR.

Not surprisingly, given this context of increasing transparency and behavioral tracking, a new field of surveillance studies has evolved. In one of the early and formative works of this interdisciplinary field, Andrejevic (2009) introduced the notion of a continuously expanding digital enclosure, in which heavily annotated information on everything people around the world did in their daily lives could be cross-referenced and searched from all angles. Andrejevic noted that more and more "analog" activities were brought into the digital enclosure each year, continuously minimizing the amounts of public and private spaces that were not monitored, recorded and processed. Solove (2004, 2007, 2011) developed a sophisticated critique of the development of privacy in the context of increasing transparency of personal data. Most notable was his deconstruction of the "I have nothing to hide"argument, which has been the common gut reaction from users when routinely waving what was earlier seen as fundamental rights (Schneier, 2006). Solove (2011) identified weaker (I have nothing to hide so others, not me, should be scrutinized) and stronger (I'm willing to sacrifice privacy for national security) variations of this argument. The digital enclosure available for Google increased dramatically as, in 2012, it created a common privacy policy for some 50 services. This enabled Google to view data streams from sources such as YouTube, Google search engine, Google apps, Google Earth and Google analytics as one single enclosure of data. Google has argued that this allows users increased privacy control functionality (http://privacy.google.com/). However, user tools available are limited compared to what has been suggested by proponents of privacy-enhancing identity management technology (Hansen, et al., 2004). In addition, van Dijck (2013) points out that it is deceptive to infer to users that they can control sharing of personal information, as controls only apply to the front-end (network of friends) and not to the back-end (third parties who purchase data).

A central issue involved in ethical discussions is whether Personally Identifiable Information (PII) can be successfully anonymized. This is basically a technological discussion surrounding the ubiquitous possibility of re-identification of once anonymized data (Ohm,

2010). Confronted with this problem, Google employee Hemerly (2013: 30) suggests injection of "additional noise" in order to make re-identification more difficult. From that perspective, the value of data-driven innovation is so substantial that privacy concerns should be taken into account but not be allowed to block progress. Hansen et al. (2008) suggest several features to strengthen privacy-oriented identity management. These include minimal disclosure tokens (utilizing cryptographic software to create multiple private certificates of identity), machine-readable privacy policies (forcing corporations to uphold privacy policies beyond the bare baseline for informed consent), sticky policies (tagging data with relevant privacy policy to avoid misuse by third parties) and transparency tools (historical data track tools that allow users to see what information has been disclosed to whom). It should also be borne in mind that privacy is not only a matter of personal values and preferences, but also may involve a contextual element where the trade-off between the risks of disclosure and the benefits of a service can be moderated for example by usage experience and trust (Martin & Shilton, 2016). Vitak, Shilton & Ashktorab (2016) further address how ready access to online data is presenting researchers with new ethical challenges of respect, beneficence, and justice.

Data work at a meso level

The previous section has highlighted the structural conditions at a macro-level pertaining to two key issues involved in negotiating a trade-off between the value and risks of data work at a micro-level level. The following section highlights how organizational initiatives in information and data governance have attempted to address the balance between the value and risks of data work at a meso-level[1]. The term 'governance' has Latin and Greek roots which associate it with the idea of 'steering'. The "idea of steersman - the person at the helm - is a particularly helpful insight into the reality of governance" (Tricker, 1984: 9). Governance can be further defined as "the fact that (a person, etc.) governs; the action or manner of governing; and the state of being governed" (Onions, 1973: 874). This definition further expands the concept and means that an adequate understanding of governance includes at least the following concerns: Who is governing? How is governance accomplished and via which mechanisms, e.g. legal and other regulations, policy-making, professional guidelines, decision-making procedures? To what is governance being directed, e.g. data, information, documents, and electronic records. Therefore a full appreciation of the emergence of data as an informational resource, and its negotiating the status of this resource as a source of value and risk, requires attention not only to the macro-context but also to the organizational conditions at a meso-level impinging on data work at a micro-level (Foster, 2016). Understandings of information governance have largely emerged out of the literatures relating to corporate governance, IT governance, information systems, and latterly information studies, while understandings of data governance have emerged more from business computing and computing science. The literatures also have different emphases. Whereas information governance focuses squarely on the trade-off between value and risk, data governance tends to look more to accountability issues, structural responsibilities and decision-making capacities, in response to external regulatory pressures as well organizational goals.

[1] Attention is restricted here to organizational issues. It is important to bear in mind however that similar issues of governance and accountability will be worked out in the extra-organizational context of consortia and the like (see for example Cutcher-Gershenfeld, J., et al. (2017).

Information governance context: enabling value and mitigating risk

Smallwood (2014: 6) refers to information governance as a "rather new multidisciplinary field that is still being defined." Definitions or descriptions that have been put forward vary according to the perspective (academic or professional) and/or discipline of their authors (e.g. records and information management, risk management and audit, law, information technology, information security). However, a thread can be discerned in terms of how the field of information governance points to the trade-off between value and risk, and its dual function of supporting the derivation of value, and mitigation of risks of working with data. Research and advisory firm Gartner define information governance as:

> "the specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival [sic] and deletion of information. It includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals" (Logan, 2010).

Other definitions in the professional domain focus on the processes or activities of information governance and also highlight its duality of purpose. For example, Hulme (2012: 100) share's IBM's perspective as it being:

> "a holistic approach to managing and using information for business benefits that encompasses information quality, information life-cycle management, and security, privacy and compliance".

This focus on business and values is perhaps most pronounced in the Information Governance Initiative (IGI) definition of information governance as:

> "the activities and technologies that organizations employ to maximize the value of their information while minimizing associated risks and costs" (IGI, 2014: 2 & 12).

Risk, in the context of compliance, has primacy, whereas in the context of data work it is value that has primacy. Hence, rather than perceiving information governance as defensive and burdensome, pertaining to compliance, it is positioned as having tangible business benefits. This duality is also exemplified by Tallon, Short and Harkins (2013) in their description of the evolution of information governance at the Intel Corporation from an era of 'protect' to one of 'protect-to-enable'.

Focusing on how information governance can add value Kooper, Maes and Roos Lindgreen (2011) appear to be the only academics to have proposed a research agenda for information governance, one that is aimed at being academically rigorous and practically relevant, to explore the optimization of information value and the roles of the stakeholders (creators, receivers and governing actors). Drawing on the concepts of governance, corporate governance, IT/ICT governance, and data governance, they offer a comprehensive definition of information governance as:

> "establishing an environment and opportunities, rules and decision-making rights for the valuation, creation collection, analysis, destruction, storage, use and control of information; it answers the question 'what information do we need, how do we make use of it and who is responsible for it?'" (Kooper, Maes & Roos Lindgreen, 2011: 195).

Tallon, Ramirez and Short (2013) cite their work, along with that of Weber, Otto and Osterle (2009) and Khatri and Brown (2010) in the data governance context, in offering their definition of information governance as:

> "a collection of capabilities or practices for the creation, capture, valuation, storage, usage, control, access, archival [sic], and deletion of information over its life cycle" (Tallon, Ramirez and Short, 2013 p. 142).

Tallon, Ramirez and Short's (2013) information governance model comprises three elements – structural practices (policy, oversight and responsibilities), procedural practices (information classification, access, data protection and backup, retention and storage migration, cost), and relational practices (user education, communication). Based on an extensive review of the literature and validated by a series of interviews with senior IT executives, their model (deriving from IT governance) also includes a series of antecedents and consequences. The former can enable or inhibit delivery of information governance. The latter represent the effects of its delivery in terms of organizational performance and risk mitigation. Again this reflects the duality identified in the definitions earlier.

Hence, whilst there is currently no commonly agreed definition of information governance, either in the professional or the academic domains, it is clear that common themes and vocabulary are emerging in their literatures, which could lead to consensus. This consensus about the definition, scope and role of information governance appears then to be developing around three key elements: (1) Who is responsible, i.e. who holds the decision rights and who is accountable? (2) How is it carried out e.g. via policy, procedures, processes and standards for 'life cycle' management of information; and (3) To what is it being directed, i.e. 'value' to the organization and its stakeholders, which encompasses value in the sense of compliance and risk management. In addition, the emerging emphasis on the duality of information governance serves to support it being a mechanism for balancing the trade-off between the value and risks of data work.

To operationalise this effectively in practice, frameworks or models are needed; and a range of frameworks have emerged which, to a greater or lesser extent, reflect the who, how and to what of information governance. These include: the Information Governance Toolkit developed by the UK National Health Service (NHS) (HSCIC Health & Social Care Information Centre, Department of Health https://www.igt.hscic.gov.uk/) and underpinned by national legislation, records management standards and sector specific requirements, in particular the Caldicott Principles which protect patient identity (Caldicott Review, 2013); ARMA International's Information Governance Maturity Model (IGMM), based on their Generally Accepted Recordkeeping Principles (GARP) (ARMA, 2009); The Information Governance Reference Model (IGRM) (EDRM.net, 2012), and a framework from the Information Governance Institute (IGI, 2014).

Information governance has been adopted as a mandate or priority for a number of professions, including records managers, health information managers, risk managers, and security specialists. Their respective professional literatures reflect significant coverage over the last decade, ranging from awareness raising, discussion of issues and regulatory/legislative compliance, to 'rally cries' for action and leadership and, more latterly, implementation strategies. It has also been covered in a number of different sectors, with the health sector being the most prominent one to feature in the literature on information governance practice (Donaldson and Walker, 2004). The extensive professional literature in the health sector is complemented by some academic research such as Gillies (2015), Liaw et al (2014), Renaud (2014), Hovenga (2013) and Hillard (2011). Research in other contexts includes that of Lajara and Macada (2013) in data quality, de Abreu Faria, Macada and Kumar (2013) in banking, Silic and Back (2013) in relation to mobile devices, and Rolfe (2015) in relation to enterprise social software.

Data governance: value and accountability
Taking information assets (or data) as "facts having value or potential value that are documented" (Khatri and Brown, 2010: 148), Khatri and Brown define data governance as "who holds the decision rights and is held accountable for an organization's decision-making about its data assets" (Khatri and Brown, 2010: 149). Data governance therefore entails initial identification of who is responsible and accountable for data assets, and the structural roles and responsibilities or loci of accountability of those who realize value from them. These roles include a range of data and information professionals including data owner/trustee, data quality manager, enterprise data architect/data modeller, data steward, data producer/supplier, data consumer, data security officer, and information chain manager. As the roles imply, there is a clear emphasis on how different roles aid in the realization of value from data in ways that are also accountable.

How will data governance be approached? As a design framework Khatri and Brown suggest that each aspect of data governance constitutes a decision domain, and that the "assignment of the locus of accountability for each decision domain will be somewhere on a continuum between centralized and decentralized" (Khatri and Brown, 2010: 151). The main mechanism proposed for practising data governance is to draw attention to processes of decision-making; and that decision-making actions can take place on a continuum from centralized to decentralized. For example, where data principles are concerned, locus of accountability can rest with a group of corporate executives, while responsibility and decision rights for data quality may rest with business division or unit managers, and therefore be decentralized. In turn, data and information professionals may be concerned with decision-making in relation to accountable data access and an accountable data life cycle. While pointing to decision-making as the main mechanism by which data is governed, Khatri and Brown (2010) also point to other structural (e.g. standing committees) and non-structural mechanisms (e.g. consistent processes, corporate announcements via web-based portals), via which data is governed, and its value realized.

What is data governance directed at? Inheriting their understanding from previous approaches to IT governance (Peterson, 2004; Weill & Ross, 2004, 2005), Khatri and Brown suggest that there are five domains that data governance shapes. These domains are data principles "clarifying the role of data as an asset"; data quality "establishing the requirements of intended use of data"; metadata "establishing the semantics or content of data so that it is interpretable by the users"; data access "specifying access requirements of data"; and the data lifecycle "determining the definition, production, retention and retirement of data" (Khatri & Brown, 2010: 149). In practice, Otto (2011) presents evidence from a comparative case study in the telecommunications sector. The key finding is that data governance is a matter not only for data and informational professionals but also other professionals, while its implementation is also contingent on organizational context. This echoes Weber, Otto and Osterle's (2009) assertion that 'one size does not fit all'.

Data work at a micro level: the roles of data, information, and other professionals

Having drawn attention to the structural conditions that exist at the macro and meso levels to shape data work, its value and risks, at a micro level we turn attention to the interactive context of data, information, and other professionals within which data work processes are embedded.

A number of professional groups play a direct and indirect role in the micro context relevant to data work. These include workers directly involved in the process of realizing value from data, including data scientists, data architects, data analysts, data visualization experts and other data professionals; along with a range of end-users who make decisions and take actions on the basis of data. To these groups, can be added those with a more supervisory and strategic interest in data work, including records managers, information governance managers and officers; data protection managers, IT risk and governance managers, and IT security personnel. Interest in realizing value from data is not restricted to data and information professionals but also includes end-users and other groups. Depending on the sector and type of organization these end-users will include business users, human resource and performance managers, financial analysts, doctors and medical administrators etc. to whom can be added a range of other interest groups outside the organization. These include: government representatives, policy-makers, independent authorities e.g. Information Commissioners' Offices, advocacy groups; as well as bodies relevant to a specific sector and who represent citizens, patients, consumers etc.

In this way we can see how the interests of a range of groups of people - and the interactions between then (see for example, Tallon, Ramirez and Short (2013), Kooper, Maes and Roos Lindgreen, 2011, edrm.net) - are of direct and indirect relevance to the question of data work. While each of these groups will play its own specialist role in processes of data work, their organization, governance, and regulation we pick out one issue that is of current relevance: the education of data scientists. While there is no doubt that the demand for data scientists is currently increasing and on an upward curve (Misnevs and Yatskiv, 2016; e-Skills UK, 2013), the supply of knowledge and skills tends towards training rather than education. Therefore, there is a need to shift, in educational terms, from the training of the data scientist to the education of the data professional. For example, several frameworks for the education of data professionals have been developed, but mainly with a focus on the data scientist (Waller and Fawcett, 2013; Granville, 2014; Song and Zhu, 2015). The EDISON project, an EU-funded project to accelerate the creation of the data science profession, has also put forward a Data Science Competence Framework (EDISON, 2016). Yet, none of these frameworks encompass the whole information life and the data value cycle: from data collection, to analyses, to interpretation to storage, all within a matrix of external conditions and consequences. Data work is largely considered in a decontextualized fashion. This article has served to show that this approach is necessary but not sufficient. There is also a need for education to develop a capacity not only for the day-to-day operations and processes of data work, but also to raise awareness of the legal, policy, ethical (see for example Fleischmann, Hui, and Wallace (2016) in the related domain of computational modelling), organizational, and governance problems and issues. In doing so, due consideration can be given to how an understanding of the context of the macro, meso and micro conditions of data work, and their consequences for organizations and individuals, can aid in the resolution and prevention of the problems that can arise in relation to the value and risks of data work.

An Integrative Framework for Future Data Work Research

It is suggested that future directions for research into data work will need to attend to the following, among other, topics:

- What values can be added to data work?
- What interactions are required to negotiate the value and risks of data work at the micro-level?
- How do structural conditions at a macro-level enable and/or constrain the effectiveness of data work at a micro-level?
- How do the consequences of data work shape its structural conditions?

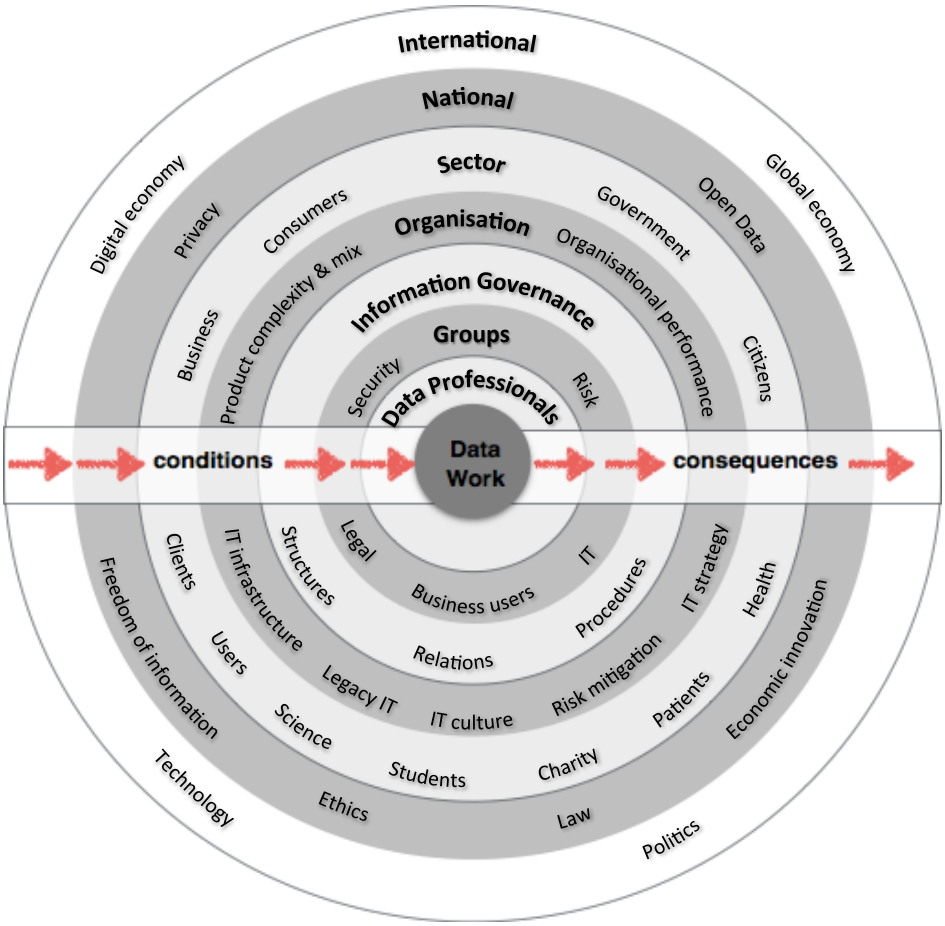

Figure 1. Data Work in Context (Foster, 2016)

To this end, an integrative framework is presented for studying these and other questions; along with the relations between data work, its processes, its antecedent structural conditions and possible consequences for individuals, organizations, and society. Figure 1: *Data Work in Context* (Foster, 2016; Corbin and Strauss, 2015) illustrates the relevant contexts. At the core of the diagram sits data work. The arrows pointing in the direction of data work indicate the antecedent conditions at a macro, meso and micro level *conditions* that influence its emergence. The arrows pointing in the direction away from data work indicate its *consequences*, for individuals, organisations, and society, at micro, meso and macro levels.

Beginning at the outer edge of the diagram, the broadest context is that of international, national, and sector conditions. International conditions will include the global economy, networked and mobile technologies; as well as relevant international relations and flows, e.g. the training and mobility of an international workforce, data flows etc. National conditions include economic and business innovation, as well legal regulations —in turn underpinned by ethical concerns —that are national in scope. Sectors are the differentiated economic sectors of work where these international and national conditions begin to have practical relevance e.g. health sector, business sector. It is at this level, for example, that sector wide regulations such as the Sarbannes-Oxley Act 2002 begin to become relevant. Exploration of these international, national, and sector conditions goes some way to explaining why data work has emerged, i.e. for its economic value, but also within a needed context of regulation and ethics. The next two meso-level contexts identify the organisational and information governance conditions relevant to enabling the value and mitigating the risks of data work. Organisational conditions include an organisation's IT capability, including IT strategy, infrastructure, legacy systems, and culture; along with the complexity of the organisation's goods and services. Corporate governance mechanisms will also seek to balance the interests of internal and external stakeholders, for example in relation to the economic value of data, and the reputational risks of any data breach. Information governance and data governance are further mechanisms that seek to exploit the value of data while also mitigating its risks. While information governance seeks control, of information assets actually created, e.g. digital records, data governance provides a mechanism for decision-making around data as an asset and how data can be turned into information. The micro context of data work identifies the structural, procedural and relational conditions, actions and interactions between professionals relevant to how data work gets done. These groups include IT professionals, legal specialists, risk and security professionals, health and business users; along with data and information professionals. The increasing interest in the value of data as both an economic asset and a societal good, and its conditions, will have consequences at other levels of the diagram. For example increased data work will have an influence on the recruitment of data professionals; the emergence of data-intensive organisations will have consequences for citizens, consumers, patients etc. In turn there will be an impact on legal regulations and ethical considerations. Beyond this the contribution of data-driven products and services will have an impact on the size and nature of the digital economy. This brings us to a final important principle of the diagram, that data work and its processes are not only conditioned by and have consequences for individuals, organisations, and economies; but these consequences also influence the initial conditions. Therefore, we can speak of a set of interrelated conditions and consequences at macro, meso and micro levels. The article has discussed these interrelated conditions and consequences in more detail without apportioning weight to any specific level or layer, save to maintain the distinction between the proximal interactional context within which data work is embedded, and the more distal structural conditions that shape this interaction.

Conclusion

The received assumption has been that data work consists of a largely unspecified set of processes implemented within an operational context separate from the matrix of conditions and consequences within which it can be located. Where attention has been given to the context of data work, this has largely been at a macro level, with less attention accorded to its conduct at meso and micro levels. The education of data scientists being a case in point. This article responds to this situation by i) systematizing data work as a set of organizing, analysing and judging, and decision-making processes, and ii) setting data work and its processes within the context of the conditions and consequences that have already and will

continue to shape the course of its future development. As data work further develops as an organised set of processes which are integral to societal activities of all kinds, it can be expected that micro-level interactions around its value and risks will entail consequences that will in turn influence its conditions. In doing so attention is drawn not only to data work as a form of organized work, but also to the 'work' required to locate data work within the sociological context of its antecedent conditions and consequences.

References

Andrejevic, M. (2009). *iSpy: Surveillance and power in the interactive era*. Kansas, US: University Press of Kansas.

ARMA International. (2009). *Generally accepted recordkeeping principles* (GARP). Available from https://www.arma.org/

Bachner, J., Ginsberg, B. & Hill, K.W. (Eds.) (2017). *Analytics, policy, and governance*. London: Yale University Press.

Borgman, C.L. (2015). *Big data, little data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Bygrave, L. A. (2002). The technologisation of copyright: Implications for privacy and related interests. *European Intellectual Property Review*, *24*(2), 51-57.

Caldicott Review, F. (2013). *Information: to share or not to share*? *The information governance review*, https://www.gov.uk/government/publications/the-information-governance-review

Childs, S and McLeod, J (2015). A case example of public trust in online records – The UK care.data programme. Final Report. InterPARES Trust Project EU17. https://interparestrust.org/assets/public/dissemination/EU17_20150802_UKCareDataProgramme_FinalReport_Final.pdf

Corbin, J. and Strauss, A. (2015). *Basics of qualitative research*: *Techniques and procedures for developing grounded theory*. 4th edition. London: Sage.

Costa, F.F. (2014). Big data in biomedicine, *Drug Discovery Today*, *19*(4), 433-440.

Crews, K.D., (1998). Harmonization and the goals of copyright: property rights or cultural progress. Ind. J. Global Legal Stud., 6 (1) [online]. Available from: http://www.repository.law.indiana.edu/ijgls/vol6/iss1/4/

Cutcher-Gershenfeld, J., Baker, K.S., Berente, N., Flint, C., Gershenfeld, G., Grant, B., … Zaslavsky, I. (2017). Five ways consortia can catalyse open science. *Nature*, *543*(7647): 615-617. http://doi.org/10.1038/543615a

Davenport, T.H. & Harris, J.G. (2017). *Competing on analytics*: *The new science of winning.* Boston (Ma.): Harvard Business Review Press.

de Abreu Faria, F., Macada, A.C.G. & Kumar, K. (2013). Information governance in the banking industry. *IEEE 46th Hawaii International Conference on System Sciences*, 4436-45. DOI 10.1109/HICSS.2013.270

DMCA (1998). *Digital Millenium Copyright Act* [online]. Available from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=105_cong_public_laws&docid=f:publ304.105.pdf

Dijck, J. van, (2013). *The culture of connectivity: A critical history of social media*. Oxford: Oxford University Press.

Donaldson, A. and Walker, P. (2004). Information governance—a view from the NHS. *International Journal of Medical Informatics*, *73*, 281-284.

Dormehl, L. (2015). *The formula*: *How algorithms solve all our problems … and create more*. London: WH Allen.

Edelstein, P. (2013). Emerging directions in analytics: Predictive analytics will play an indispensable role in healthcare transformation and reform. *Health Management Technology*, *34*(1), 16-17.

EDISON (2017). *Data science competence framework.* Available from http://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf

EDRM.net. (2012). *Information governance reference model* (IGRM). Available from https://www.edrm.net/frameworks-and-standards/information-governance-reference-model/

e-skills UK (2013). *Big data analytics*: *An assessment of demand for labour and skills*, 2012-2017. Available from https://www.thetechpartnership.com/globalassets/pdfs/research-2013/bigdataanalytics_report_jan2013.pdf

European Parliament and Council (2016). *General Data Protection Regulation*. *L119*, 4/5/2016, 1–88. Available from European Parliament and Council (2016). *General Data Protection Regulation*. *L119*, 4/5/2016, 1–88

Fleischmann, K.R., Hui, C. & Wallace, W.A. (2017). The societal responsibilities of computational modellers: Human values and professional codes of ethics. *Journal of the Association for Information Science and Technology*, *68*(3), 543-552.

Foster, J. and Clough, P.D. (2018). Embedded, added, co-created: Re-visiting the value of information in an age of data, *Journal of the Association for Information Science and Technology*, *69*(5), 744-748.

Foster, J. (2016). "Towards an understanding of data work in context: Emerging issues of economy, governance, and ethics", *Library Hi Tech*, *34*(2), 182-196.

Furner, J. (2016). "Data": The data. In M Kelly & J Bielby (Eds.). *Information cultures in the digital age: A Festschrift in honor of Rafael Capurro* (pp.287-306). Springer. http://doi.org/10.1007/978-3-658-14681-8_17

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Gillies, A. (2015). The role of information governance within English clinical governance: Observations based upon the interim report from the NIGC of the Care Quality Commission. *Clinical Governance*: *An International Journal*, *20*(1), 13-20.

Gillespie, T. (2007). *Wired shut*. Boston MA: MIT Press.

Granville, V. (2014). *Developing analytic talent: Becoming a data scientist.* Indianapolis (IN): John Wiley.

Greenberg, J., Ogletree, A, Murillo. A.P., Caruso, T.P., and Huang, H. (2014). Metadata capital: Simulating the predictive value of self-generated health information (SGHI), *IEEE International Conference on Big Data*, 31-36.

Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA, and the US surveillance state*. Basingstoke: Macmillan.

Hansen, M., Berlich, P., Camenisch, J., Clauß, S., Pfitzmann, A., & Waidner, M. (2004). Privacy-enhancing identity management. *Information Security Technical Report*, *9*(1), 35-44.

Hansen, M., Schwartz, A., & Cooper, A. (2008). Privacy and identity management. *IEEE Security & Privacy*, *6*(2), 38-45.

Hemerly, J. (2012). Policy Colloquium Notes: Empowering Data-Driven Innovation-Co-convened by Google and O'Reilly Media at Google's Mountain View Campus. Available at SSRN 2087794.

Hemerly, J. (2013). Public policy considerations for data-driven innovation. *Computer*, *46*(6), 25-31.

Hustinx, P. (2010). Privacy by design: delivering the promises. *Identity in the Information Society*, *3*(2), 253-255.

Hovenga, E.J.S. (2013). National healthcare systems and the need for health information governance. *Studies in Health Technology and Informatics*, *193*(3), 3-23.

IGI (2014). *Annual Report 2014: Information governance goes to work*. Available from http://www.iginitiative.com/

Janke, A.T., Daniel, BS, Overbeek, D.L., Kocher, K.E., & Levy, P.D. (2016). Exploring the potential of predictive analytics and big data in emergency care. *Annals of Emergency Medicine*, *67*(2), 227-236.

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*, 2561-2573.

Kooper, MN, Maes, R & Roos Lindgreen, EEO. (2011). On the governance of information: Introducing a new concept of governance to support the management of information. *International Journal of Information Management*, *31*, 195-200.

Kauffman, R.J., Srivastava, J., and Vayghan, J. (2012). Business and data analytics: New innovations for the management of e-commerce. *Electronic Commerce Research and Applications*, 11, 85-88.

Kayyali, B., Knott, D. & Van Kuiken (2013). The big-data revolution in US health care: Accelerating value and innovation. McKinsey & Company. Available from: mckinsey.com

Khatri, V. K. & Brown, C.V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148-152.

Kooper, M.N., Maes, R., & Roos Lindgreen, E.E.O. (2011). On the governance of information: Introducing a new concept of governance to support the management of information. *International Journal of Information Management*, *31*, 195-200.

Kwon, O., Lee N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387-394.

Lajara, T.T. & Macada, A.C.G. (2013). Information governance framework: The defense manufacturing case study. *Proceedings of the Americas Conference on Information Systems*, 3, 1984-1993.

Lane, J, Stodden, V. Bender, S., & Nissenbaum, H. (2014). *Privacy, big data, and the public good*: *Frameworks for engagement*. Cambridge: Cambridge University Press.

Laney, D. (2001). *3-D Data Management: Controlling Data Volume, Velocity, and Variety*. Available from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larsson, S., & Svensson, M. (2010). Compliance or obscurity? Online anonymity as a consequence of fighting unauthorised file‐sharing. *Policy & Internet*, *2*(4), 77-105.

Liaw, S., Pearce, C., Liyanage, H., Liaw, G.S.S. & de Lusignan, S. 2014. An integrated organisation-wide data quality management and information governance framework: theoretical underpinnings. *Informatics in Primary Care*, *21*(4), 199-206.

Litman, J. (2001). *Digital copyright*. Prometheus books.

Logan, D. (2010). What is information governance and why is it so hard? Available from http://blogs.gartner.com/debra_logan/2010/01/11/what-is-information-governance-and-why-is-it-so-hard/

Martin, K. & Shilton, K. (2015). Why experience matters to privacy: How context-based experience moderates consumer privacy expectations for mobile applications, *Journal of the American Society for Information Science and Technology*, *67*(8), 1871-1882.

May, C., (2003). Digital rights management and the breakdown of social norms. *First Monday*, *8* (11). Available from http://firstmonday.org/issues/issue8_11/may/index.html

McLeod, J & Childs S. (forthcoming 2018). Public trust in online records: The case of the UK care.data programme. In: Anderson, A, Becker, IC & Duranti, L. (Eds). Born Digital in the Cloud: Challenges and Solutions, Presentations at the 21st Archival Sciences Colloquium of the Marburg Archives School, p. 43-64. ISBN: 978-3-923833-83-2

Miller, H.G. & Mork, P. (2013). From data to decisions: a Value chain for big data. *IT Professional*, 57-59.

Misnevs, B. and Yatskiv, I. (2016). Data science: Professional requirements and competence evaluation. *Baltic Journal of Modern Computing*, *4*(3), 441-453.

Neff, G. (2013). Why big data won't cure us, *Big Data*, *1*(3), 117-123.

OECD (1980). *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data: Explanatory Memorandum*. Paris: OECD.

OECD (2011). *The Evolving Privacy Landscape: 30 years after the OECD privacy guidelines*. Paris: OECD.

OECD (2015). *Data-Driven Innovation: Big Data for Growth and Well-Being*. Paris: OECD Publishing.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701.

Onions, C.T. (1973). *The Shorter Oxford English Dictionary on Historical Principles, Volume 1*: *A – Markworthy*. Oxford: Clarendon Press.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies*, *65*, 17-37.

Otto, B. (2011). Organizing data governance: Findings from the Telecommunications Industry and consequences for large service providers. *Communications of the Association for Information Systems*, *29*(3), 45-66.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135.

Peterson, R. (2004). Crafting information technology governance. *Information Systems Management*, *21*(4), 7-22.

Ragan, CR. (2013). Information governance: it's a duty and it's smart business. *Richmond Journal of Law & Technology*, XIX (4), pp. 1-50. http://jolt.richmond.edu/v19i4/article12.pdf

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(3), 1-10.

Renaud, K. (2014). Clinical and information governance proposes; human fallibility disposes. *Clinical Governance: An International Journal*, 19 (2), 94-109.

Rolfe, O. (2015). Information governance for enterprise social software. *Business Information Review*, 32(1), 38-44.

Schneier, B. (2006). The eternal value of privacy. Comment on Wired. com, May.

Schroeder, A. & Wagner, C. (2012). Governance of open content creation: A conceptualization and analysis of control and guiding mechanisms in the open content domain. *Journal of the American Society for Information Science and Technology*, *63* (10), 1947-1959.

Sell, S., & May, C. (2001). Moments in law: contestation and settlement in the history of intellectual property. *Review of International Political Economy*, *8*(3), 467-500.

Shilton, K. (2012). Participatory personal data: An emerging research challenge for the information sciences. *Journal of the American Society for Information Science and Technology*, *63*(10), 1905-1915.

Silic, M. & Back, A. (2013). Factors impacting information governance in the mobile device dual-use context. *Records Management Journal*, *23*(2), 73-89.

Smallwood, R.F. (2014). *Information governance*: *concepts, strategies, and best practices.* London: Wiley.

Smith, S., (2003). From Napster to Kazaa: The battle over peer-to-peer filesharing goes international. *Duke Law & Technology Review*, *2*(1), 1-9.

Solove, D. J. (2004). *The digital person: Technology and privacy in the information age*. NyU Press.

Solove, D. J. (2007). 'I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review*, *44*, 745. Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=998565

Solove, D. J. (2011). *Nothing to hide: The false tradeoff between privacy and security*. Yale University Press.

Song, I-Y. and Zhu, Y. (2015). Big data and data science: What should we teach? *Expert Systems*, **33** (4), 364-373.

Sterne, J. (2006). The mp3 as cultural artifact. *New Media & Society*, *8*(5), 825-842.

Stein, A. (2012). Big data and analytics, the analytics value chain – part 3. Available at http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

Tallon, P.P., Ramirez, R.V., & Short, J.E. (2013). The Information artifact in IT governance: Toward a theory of information governance, *Journal of Management Information Systems*, *30*(3), 145-181.

Tallon, PP., Short, JE. & Harkins, MW. (2013). The evolution of information governance at Intel. *MIS Quarterly Executive*, *12*(4), p. 189-199.

Taylor, R.S. (1986). *Value-added processes in information systems*. Norwood (NJ): Ablex Publishing Corporation.

Tene, O. & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics., *Northwestern Journal of Technology and Intellectual Property*, 239. Available at: SSRN: https://ssrn.com/abstract=2149364

Tricker, R. (1984). *Corporate governance, practices, procedures and powers in British companies and their boards of directors*. Oxford: The Corporate Policy Group.

Turow, J. (2012). *The daily you: How the new advertising industry is defining your identity and your worth*. Yale: Yale university press.

US Copyright Act., (1976). [online]. Available from: http://www.law.cornell.edu/copyright/copyright.table.html

US Congress (1998). Digital millennium copyright act. *Public Law*, *105*(304), 112.

Vitak, J., Shilton, K. & Ashktorab, Z. (2016). Beyond the Belmont Principles: Ethical challenges, practices, and beliefs in the online data research community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing – CSCW'16*. ACM Press. http://doi.org/10.1145/2818048.2820078

Waller, M.A. and Fawcett, S.E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management*, Journal of Business Logistics*, *34*(2), 77-84.

Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all---A contingency approach to data governance. *Journal of Data and Information Quality*, *1*(1), Article 4 (June 2009), 27 pages. DOI: 10.1145/1515693.1515696

Weill, P. & Ross, J.W. (2004). *IT governance: how top performers manage IT decision rights for superior results. Boston* (Ma.): Harvard Business School Press.

Weill, P., & Ross, J.W. (2005). A matrixed approach to IT governance, *MIT Sloan Management Review*, Winter 2005, *46*(2), 26.

Wilbanks, J. & Friend, S.H. (2016). First, design for data sharing. *Nature Biotechnology*, *34*, 377-379.

WTO (1994). *Agreement on Trade Related Aspects of Intellectual Property Rights* [online]. Available from: http://www.wipo.int/wipolex/en/other_treaties/details.jsp?group_id=22&treaty_id=231